MsDetector

MsDetector is designed and developed by Hani Zakaria Girgis under the supervision of Sergey Sheetlin at the Spouge Research Group, the National Center for Biotechnology Information, The National Institutes of Health, USA.

The library HMMlib is utilized in MsDetector. HMMlib is covered by the GNU Lesser General Public License.

Version

1.2

Contact

If you have questions, feel free to contact the first author at girgishz@mail.nih.gov or hani.z.girgis@gmail.com

MsDetector with Optimized Parameters

Usage

```
MsDetectorOptimized[32/64/Mac] sequence_file masked_sequence_file ms_file
```

Input

The sequence_file is the input file in FASTA format. Any valid FASTA header is accepted. If the user wishes to use a specific offset/start, the header form is >chromosome:start-end. The start in the header refers to the first nucleotide in the sequence. In this case, the locations of the found microsatellites (MSs) are relative to the start the user provided. MsDetector can process multiple sequences in the same input file.

Example input:

```
>chrX:0-36
ATATATATATATATATATATATATATATATATATAT

>gi|224514821|ref|NT_167199.1| Homo sapiens chromosome Y genomic contig, GRCh37.p5 Primary Assembly
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTCTGAAAGTGGACCTATCAGCAGGATGTGGGTG
GGAGCAGATTAGAGAATAAAAGCAGACTGCCTGAGCCAGCAGTGGCAACCCAATGGGGTCCCTTTCCATA
CTGTGGAAGCTTCGTTCTTTCACTCTTTGCAATAAATCTTGCTATTGCTCACTCTTTGGGTCCACACTGC
CTTTATGAGCTGTGACACTCACCGCAAAGGTCTGCAGCTTCACTCCTGAGCCAGTGAGACCACAACCCCA
CCAGAAAGAAGAAACTCAGAACACATCTGAACATCAGAAGAAACAAACTCCGGACGCGCCACCTTTAAGA
```

Output

If any of the two output files is present in the current directory, MsDetector deletes it before the program starts. The masked_sequence_file contains the masked sequence(s). Microsatellites are written in lower case letters. The ms_file contains the locations of the microsatellites and their scores, which are greater than or equal 0.5 and less than 1.0.

The coordinates system

Coordinates are zero-based. The start coordinate is inclusive, whereas the end coordinate is exclusive. Consequently, length = end - start. Use this convention to specify a header of the form >chromosome:start-end. The locations outputted by MsDetector follow this convention as well.

MsDetector with User-specified Parameters

Overview of the procedure implemented in MsDetector

1. Score sequence(s),
2. Detect potential MSs, and
3. Apply the filter.

Parameters

Mandatory parameters:

```
-seq sequence_file
-len motif_length
-fct win_factor [window = 2 x win_factor x motif_length]
-mtr matrix_name [Id, Trans, Comp, and TransComp]
```

The input sequence file can be specified with the parameter -seq. Motif length can be provided through the -len parameter. The length of one of the two flanking sequences, i.e. a half window, is a multiple of the motif length. The multiplication factor must be provided with the -fct parameter. For example, if the motif length is 6, a factor of 4 results in a half window of size 24 nucleotides.

The name of the scoring matrix can be provided to MsDetector via the -mtr parameter. Our experiments show that the identity matrix, Id, results in the most efficient and accurate detections. In the transition matrix, Trans, a match between two different Purines (A & G) or two different Pyrimidines (C & T) is assigned half of the score assigned to a match between two identical nucleotides. The transition matrix leads to a comparable performance, but it slows the execution. The composition-correction matrix, Comp, is designed for genomes with unusual composition, e.g. AT-rich genomes. The scores in this matrix are based on the frequencies of the nucleotides in the genome. In general, a match between two abundant nucleotides is assigned a score less than that of two less abundant nucleotides. The composition-correction matrix, and the matrix that combines transition with composition correction, TransComp, do not lead to improvement; however, we included them for experimentation purposes.

If the user wishes to use the Comp matrix or the TransComp matrix, a file including the nucleotides probabilities in the genome of interest must be provided via the optional parameter –frq, see below. The first line of the file must include the probabilities of A, C, G, and T, in this specific order, separated by a space(s). These probabilities can be calculated by the program NucleotideFreqMaker[32/64/Mac]. This program requires a directory including the chromosomes comprising the genome of the species of interest. The sequences of the chromosomes must be in FASTA format and the files must have '.fa' extension. Each chromosome file must include one sequence only.

Optional Parameters:

```
-hmm hmm_file
-glm glm_file
-thr glm_threshold [>= 0.0 && <= 1.0]
-frq frequencies_file [must be provided with the Comp or the TransComp matrixes]
-sfl shuffle [0 or 1 default is 0]
-scr scores_file
-msk masked_sequence_file
-rpt ms_locations_file
```

The system is very flexible. Once the user provides the values of the mandatory parameters, the optional parameters can be used to control the procedure to extract MSs and to specify the output format. The user can perform any of the following four tasks by providing the values of the required parameters: (1) scoring a sequence, (2) detecting potential MSs, (3) filtering potential MSs, (4) performing any of the previous operations on a shuffled version of the input sequence.

Examples

1. Scoring a sequence

```
MsDetector[32/64/Mac] -seq seq.fa -len 6 -fct 4 -mtr Id -scr scores.sc
```

This command processes the sequence(s) in seq.fa, the motif length is 6 bp, the size of one of the two flanking sequences is 24 bp, and the identity scoring matrix is used. The scores of 500-bp-long segments are outputted to scores.sc. These segments do not include the N character, which represents an unknown nucleotide.

2. Detecting potential MSs

```
MsDetector[32/64/Mac] -seq seq.fa -len 6 -fct 4 -mtr Id -hmm hmm.txt -rpt ms.txt -msk seq_masked.fa
```

This command scores the sequence(s) and invokes the detection component of MsDetector. The locations of the potential MSs and their scores are written to ms.txt and the masked sequence(s) is written to seq_masked.fa

3. Applying the filter

```
MsDetector[32/64/Mac] -seq seq.fa -len 6 -fct 4 -mtr Id -hmm hmm.txt -glm glm.txt -rpt ms.txt
```

This command invokes the scoring, the detection, and the filtering components of MsDetector. The locations of the MSs and their scores will be written to ms.txt.

```
MsDetector[32/64/Mac] -seq seq.fa -len 6 -fct 4 -mtr Id -hmm hmm.txt -glm glm.txt -thr 0.99 -rpt ms.txt
```

This command does the same as the previous command except that the threshold of the filter is 0.99 instead of the default of 0.5. The threshold must be between zero and one inclusively.

4. Obtaining false positive detections

```
MsDetector[32/64/Mac] -seq seq.fa -len 6 -fct 4 -mtr Id -hmm hmm.txt -glm glm.txt -sfl 1 -rpt ms.txt
```

This command invokes MsDetector to shuffle the input sequence and to search for MSs in the shuffled sequence. This command can be useful in calculating the false positive rate.